DSM: Question Generation over Knowledge Base via Modeling Diverse Subgraphs with Meta-learner

Shasha Guo¹, ², Jing Zhang¹, ², Yanling Wang¹, ², Qianyi Zhang¹, ², Cuiping Li¹, ², Hong Chen¹, ²

¹School of Information, Renmin University of China, Beijing, China
²Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education
{guoshashaxing, zhang-jing, wangyanling, qianyizhang, licuiping, chong}@ruc.edu.cn

Abstract

Existing methods on knowledge base question generation (KBQG) learn a one-size-fits-all model by training together all subgraphs without distinguishing the diverse semantics of subgraphs. In this work, we show that making use of the past experience on semantically similar subgraphs can reduce the learning difficulty and promote the performance of KBQG models. To achieve this, we propose a novel approach to model diverse subgraphs with meta-learner (DSM). Specifically, we devise a graph contrastive learning-based retriever to identify semantically similar subgraphs, so that we can construct the semantics-aware learning tasks for the meta-learner to learn semantics-specific and semantics-agnostic knowledge on and across these tasks. Extensive experiments on two widely-adopted benchmarks for KBQG show that DSM derives new state-of-the-art performance and benefits the question answering tasks as a means of data augmentation. Codes and datasets are available online¹.

1 Introduction

In recent years, knowledge base question generation (KBQG) has attracted substantial research interests as it shows great promise to improve the quality of question answering (QA). Specifically, KBQG can augment training data for QA systems (Chen et al., 2020; Indurthi et al., 2017), and it can also motivate the machines to actively ask questions in human-machine conversations (Sun et al., 2018b; Zeng and Nakano, 2020).

Concretely, KBQG generates natural language questions according to a set of facts extracted from KB, where each fact is typically specified as a triplet \((e, r, e')\) meaning entity \(e\) has relation \(r\) with entity \(e'\). Previous efforts on KBQG can be categorized into template-based models (Seyler et al., 2017) and neural network-based (NN-based) models (Bi et al., 2020; Kumar et al., 2019). The former ones heavily depend on hand-crafted templates, resulting in low scalability as these templates are limited to narrow domains. Alternatively, NN-based models address this issue via inputting the set of triplets about a certain answer into a Seq2Seq architecture to automatically generate the question.

In fact, for generating a question, triplets about a certain answer can naturally form a subgraph as illustrated in Figure 1. We observe that subgraphs differ in their semantics, which is especially shown in the relations that express the triplets² as well as the structural patterns such as chain, star, and triangle³. Existing efforts do not distinguish the semantics of different subgraphs but learn a one-size-fits-all model by training together all subgraphs, which increases the learning difficulty. Inspired by humans who solve a problem by searching the relevant problems that they have encountered in the past and adjusting the solution of these problems to the new one (Lancaster and Kolodner, 1987; Ross, 1984), we avoid directly learning a model on the entire data but try to leverage the past experience from similar KBQG cases to supervise generation from the current subgraph.

To achieve the goal, we propose a KBQG approach which models diverse subgraph with meta-learner (DSM). DSM retrieves semantically similar subgraphs which share similar relations and structures to construct semantics-aware learning tasks, so that the model can carefully learn potential question generation (QG) patterns over each kind of subgraphs. With multiple learning tasks, we employ a Model-Agnostic Meta-Learning (Finn et al., 2017)-like (MAML-like) meta-learner to capture semantics-specific and semantics-agnostic knowl-

²The relations instead of the concrete entities in a subgraph is the decisive factor of the meaning.
³More complex examples can be viewed as the combination of chain, star, and triangles.

¹https://github.com/RUCKBReasoning/DSM
knowledge on and across different learning tasks.
To create the above learning tasks in DSM, retrieving similar subgraphs is crucial. Although classic graph matching algorithms (Li et al., 2019; Riba et al., 2018) can help do this, they only consider graph structural properties, regardless of the semantics of relations. Inspired by the great success of graph neural networks (GNNs) (Hamilton et al., 2017; Kipf and Welling, 2017; Velickovic et al., 2018), we turn to represent subgraphs in the embedding space by GNNs, as they can easily encode both relations and structures. By doing this, we can retrieve semantically similar subgraphs according to cosine similarity between their representations. Due to the lack of supervision, we perform graph contrastive learning (GCL) (Qiu et al., 2020; You et al., 2020a), which is one of the mainstream graph self-supervised learning methods. To enable GCL, we propose relation path-based similarity, a simple and effective metric, to retrieve similar subgraphs as positive samples of contrastive learning.

Contributions. (1) We design a KBQG approach that considers the diversity of subgraph semantics. Instead of training subgraphs of different semantics together, we construct semantics-specific learning tasks to reduce the learning difficulty. (2) We devise a GCL-based retriever to identify semantically similar subgraphs, so that we can construct semantics-aware learning tasks for the meta-learner to enable the meta-learner learn semantics-specific and semantics-agnostic knowledge. (3) Our model shows the new state-of-the-art (SOTA) performance in BLEU and ROUGE, and benefits the QA tasks as a means of data augmentation. Human evaluation and case studies also show that our model can generate more relevant and fluent questions than other baselines.

2 Related Work

Knowledge Base Question Generation. Existing KBQG models can be divided into two categories — template-based models and neural network-based (NN-based) models. The former (Seyler et al., 2017) designs heuristic templates for question generation, which is simple but has low scalability. Driven by advances of deep neural networks (Shen et al., 2018; Vaswani et al., 2017), NN-based models (Bi et al., 2020; Elsahar et al., 2018; Indurthi et al., 2017; Liu et al., 2019) are applied to generate questions automatically. Generally, triplets in a subgraph are organized into a sequence to be the input of a Seq2Seq (Sutskever et al., 2014) neural network. Since the graph topology around each entity also contains useful semantics, recent studies (Chen et al., 2020; Kumar et al., 2019) utilize graph neural networks to encode the structural patterns. Nevertheless, previous studies overlook modeling the diversity of graph semantics, which could make the model easy to get over-fitting. On the other hand, since existing NN-based models are trained from scratch, their performance thereby heavily relies on the scale of training data. Pre-trained language models (PLMs) can help solve this problem as they have been trained on the large corpus to be empowered with rich semantic information. Currently, some attempts of question generation over unstructured textual data (Chan and Fan, 2019; Dong et al., 2019) have adopted PLMs. But rare studies discuss PLM-based question gen-
eration over structured KB. To this end, we target
to design a PLM-based KBQG model which consi-
siders the diversity of subgraphs in KB.

**Graph Self-supervised Learning.** To construct semant-
ics-aware learning tasks, we need to re-
trieve subgraphs with similar semantics. This work
performs self-supervised learning (SSL) over sub-
graphs to learn their representations. By doing so,
we can retrieve semantically similar subgraphs ac-
gording to the similarity between their representations.
Graph SSL schemes, which learn node-level or graph-level representations to retain the attribu-
tive and structural patterns of the graph data, have
been widely studied in recent years. Generally,
graph SSL (Liu et al., 2021; Wang et al., 2021) can be
divided into four types, including generation-
based (Kim and Oh, 2021; You et al., 2020b), aux-
iliary property-based (Peng et al., 2020a; Sun et al.,
2020), contrastive-based (Hu et al., 2020a; Wang
et al., 2022) and hybrid (Hu et al., 2020b; Peng
et al., 2020b) methods. This paper adopts the graph
contrastive learning method to learn subgraph rep-
resentations via “graph-graph” contrast, which is more
suitable to the target of subgraph retrieval.

## 3 Preliminary

A **knowledge base** (KB) organizes the factual infor-
mation as a set of triplets, *i.e.*, \( KB = \{(e, r, e')|e, e' \in E, r \in R\} \), where \( E \) and \( R \) denote
the entity set and the relation set respec-
tively. From a \( KB \), we can extract any subgraph
\( G_i \subset KB \). For convenience, we employ \( A_i \) to
denote the adjacent matrix of \( G_i \) and use \( n_i \) to rep-
resent the number of nodes in \( G_i \).

### 3.1 KBQG

Given a set of (subgraph, answer, question)
tuples as the training data denoted by \( D = \{(G_i, a_i, q_i)\}^{N}_{i=1} \) with \( N \) as the total number of
data samples, the objective of KBQG is to learn a
mapping function \( f \) with parameter \( \theta \), i.e.,

\[
f_\theta : (G_i, a_i) \rightarrow \hat{q}_i,
\]

where \( \hat{q}_i \) denotes the predicted natural language
question consisting of a sequence of word tokens,
and \( q_i \) is the ground truth. The goal is to optimize
the model parameter \( \theta \) to maximize the conditional
likelihood \( P_{\theta}(q_i|G_i, a_i) \).

### 3.2 PLM-based KBQG Model

Inspired by the great success of pretrained language
models (PLMs), we first use BART (Lewis et al.,
2020), a pre-trained Seq2Seq model, as a base-
line to instantiate \( f_\theta \). Specifically, we linearize
each subgraph \( G_i \) into a triplet-based sequence,
where each triplet is separated by the special token
“\\langle/s\\rangle”, then we input the sequence into BART to
generate a question about the answer \( a_i^4 \).

As reported in Table 2, directly fine-tuning the
BART has already outperformed the state-of-the-
art baseline G2S+AE+RL (Chen et al., 2020). How-
ever, this straightforward solution ignores the di-
verse semantics of subgraphs. Instead of learning
a one-size-fits-all model, we model diverse seman-
tics and learn over semantically similar subgraphs,
aiming to reduce the learning difficulty.

## 4 DSM

We introduce our KBQG approach which models
Diverse Subgraph with Meta-learner (DSM).

### 4.1 Model Overview

Figure 2 illustrates the overview of our approach.
Overall, DSM contains two key components, a
subgraph retriever and a MAML-like meta-learner.
The subgraph retriever retrieves top-\( k \) similar sub-
graphs to a query subgraph to construct a learning
task. Based on multiple learning tasks, a MAML-
like meta-learner summarizes semantics-specific and semantics-agnostic knowledge on and across
these learning tasks.

Specifically, for a given query subgraph \( G_i \) of an
answer entity \( a_i \), the subgraph retriever retrieves
top-\( k \) similar subgraphs to \( G_i \), which compose the
support set \( S_i = \{(G_j, a_j, q_j)\}^{k}_{j=1} \) for the data
sample \( D_i = (G_i, a_i, q_i) \). To flexibly retrieve
the top-\( k \) subgraphs, we represent each subgraph by a
GNN encoder and leverage graph contrastive learn-
ing (GCL) to learn parameters of the GNN encoder.
A key of GCL is constructing positive sample pairs.
To enable GCL to capture the semantic similari-
ty between subgraphs, we devise a relation path-based
similarity metric to guide the positive sample pair
construction. After retrieving the support set for
each learning task, the MAML-like meta-learner
optimizes the parameter \( \theta \) of the mapping function

---

4We have empirically proved that the generated question is insensitive to the order of the triplets.
Figure 2: Illustration of the overview framework, which comprises a subgraph retriever and a MAML-like meta-learner. The subgraph retriever, implemented by a GCL-based method, retrieves top-k similar subgraphs to construct multiple learning tasks, such that the meta-learner can learn the semantics-specific knowledge on each task by the meta-train step and learn the semantics-agnostic knowledge across these tasks by the meta-test step.

More concretely, the structure is useful but not the decisive factor in the semantics of a subgraph. If relation sets of two subgraphs do not intersect, no matter how similar their structures are, the generated questions differ a lot from each other. Besides, we find that entity names can be easily copied from the input subgraph to the generated question, so it is unnecessary to consider entity names when measuring the semantic similarity.

Motivated by the great success of graph neural networks (GNNs) (Kipf and Welling, 2017; Velickovic et al., 2018), we leverage GNNs to learn a low-dimensional real-valued embedding for each subgraph, so that we can retrieve according to the cosine similarity between subgraph representations. Since GNNs are able to encode both the semantics of relations and structures, the cosine similarity between subgraph embeddings can represent the above-demanded subgraph similarity. Due to the lack of supervision, we propose to perform graph contrastive learning (GCL) (Velickovic et al., 2019; Wang et al., 2022), one of the mainstream graph self-supervised learning (SSL) methods. To enable GCL, we define relation path-based similarity, a simple and effective metric, for finding similar subgraphs as the positive sample pairs of GCL.

Next, we explain how to design a subgraph encoder for encoding both relations and structures, and how to conduct positive sample generation for contrastive learning.

4.2.1 Relation-enhanced Subgraph Encoder

We propose a relation-enhanced graph encoder for representing a subgraph, as relation information is a crucial factor in the semantics of the generated questions. For example, in Figure 1, although the entities in Figure 1(c) are totally different from
where \( r \) denotes the embedding of the relation \( r \), and \( R_j \) is the set of the relations connected to \( e_j \). Given a subgraph \( G_j \) with entities’ initial features \( H_j^{(0)} = \{ h_j^{(0)} \}_{j=1}^{n_j} \) and the adjacency matrix \( A_j \), as input, the GNN encoder \( g_\phi \) with \( L \) layers outputs the entity embeddings \( H_j^{(L)} \). Then we average all the outputted nodes’ representations and apply a sigmoid activation function on the pooled result to represent the graph-level representation \( z_i \).

Building on the relation-based initial node features, we adopt GIN (Xu et al., 2019), a SOTA GNN architecture, to instantiate the GNN encoder \( g_\phi \). In addition to the neighborhood homophily, GIN can encapsulate the structures of nodes, which is helpful for representing a subgraph.

Obviously, other heterogeneous GNN encoders such as RGCN (Schlichtkrull et al., 2018) are alternative encoders. However, most of them create a separate parameter for each relation, which ignores the relations’ natural language semantics. We will demonstrate the superiority of our method through experiments (Cf. Table 3 for details).

**Contrastive Loss Function.** We adopt the normalized temperature-scaled contrastive loss as in (Sohn, 2016; You et al., 2020a). Formally, the NT-Xent for a mini-batch is formulated as:

$$
\mathcal{L}_{GCL} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{ \exp (z_i \cdot z_i^\top / \tau) } { \sum_{j=1}^{n} \sum_{j' \neq i} \exp (z_i \cdot z_j^\top / \tau) } \, ,
$$

(3)

where \( n \) is the number of subgraphs in a mini-batch, \( m \) represents the number of positive samples for each subgraph, \( Pos(i) \) represents the indices of positive samples of the \( i \)-th sample, and \( \tau \) denotes a temperature parameter. The contrastive learning algorithm is illustrated in Algo. 1.

**Algorithm 1: Contrastive Learning**

**Input:** \( \mathcal{G} = \{ G_i \}_{i=1}^{N} \), \( m \) (positive sample size).

**Output:** \( \phi \) of the GNN encoder.

1: Initialize the parameters \( \phi \) for the GNN encoder;
2: for each epoch do
3: Sample a mini-batch of \( n \) subgraphs \( \mathcal{B} = \{ G_i \}_{i=1}^{n} \subset \mathcal{G} \);
4: for each subgraph \( G_i \in \mathcal{B} \) do
5: Generate \( m \) positive sample \( \{ G_j^+ \}_{j=1}^{m} \in \mathcal{G} \) (Algo. 2);
6: end for
7: \( \{ z_i \}_{i=1}^{n} \) = \( g_\phi (\{ G_i, \{ G_j^+ \}_{j=1}^{m} \}_{i=1}^{n}) \);
8: \( \phi \leftarrow \text{Adam} (\mathcal{L}_{GCL} (\{ z_i \}_{i=1}^{n})) \);
9: end for

**Algorithm 2: Positive Sample Generation**

**Input:** Paths of graphs \( \{ P_i \}_{i=1}^{N} \) and path-to-graph index \( \{ \mathcal{G}_p \}_{p=1}^{M} \).

**Output:** Top-\( m \) similar subgraphs about \( G_i \).

1: for each path \( p \in P_i \) do
2: for each graph \( G_j \in \mathcal{G}_p \) do
3: Add \( G_j \) into the sample candidate set \( C_i \);
4: end for
5: end for
6: for each graph \( G_j \in C_i \) do
7: Calculate \( S_{RP}(G_i, G_j) \) by Eq. (4);
8: end for
9: Return top-\( m \) similar graphs according to \( S_{RP}(G_i, G_j) \);

Intuitively, the relation path-based metric can be directly used by the subgraph retriever. However, this method fails to cover subgraphs whose relation paths are distinct but semantically similar such as \( \langle \text{Lisa},\: \text{born_in},\: \text{France} \rangle \) and \( \langle \text{Lisa},\: \text{place_of_birth},\: \text{France} \rangle \).

4.2.2 Positive Sample Generation

In this subsection, we explain how to construct positive sample pairs. The details are presented in Algo. 2. Without the ground truth, we define relation path-based similarity, a simple and effective metric for measuring the similarity between subgraphs, and propose an efficient retrieval method based on a path-to-graph index.

The relation path-based metric can be directly used by the subgraph retriever. However, this method fails to cover subgraphs whose relation paths are distinct but semantically similar such as \( \langle \text{Lisa},\: \text{born_in},\: \text{France} \rangle \) and \( \langle \text{Lisa},\: \text{place_of_birth},\: \text{France} \rangle \).
Relation Path. Given a subgraph $G_i$, a relation path is denoted by $\pi = (r_1, \cdots, r_{|\pi|})$. Any relation in $\pi$ is included in $G_i$ and is traversed following the relation direction from the arrowhead to the arrowtail. The length $|\pi|$ of path $\pi$ represents the maximal length of all possible paths in $G_i$. Figure 1 illustrates relation paths in a subgraph. For example, in Figure 1(e), we can enumerate four paths, including one $\pi_7 = \text{spouse} \rightarrow \text{live in}$, two $\pi_8 = \text{live in} \rightarrow \text{spouse}$, and one $\pi_9 = \text{spouse} \rightarrow \text{live in} \rightarrow \text{spouse}$. We use the classic DFS algorithm to obtain all paths of a subgraph.

Relation Path Similarity. Given two subgraphs $G_i$ and $G_j$, we respectively enumerate relation paths in them to construct the support set $\mathcal{P}_i$ and $\mathcal{P}_j$. Then relation path-based similarity $S_{RP}$ between $G_i$ and $G_j$ is defined as:

$$S_{RP}(G_i, G_j) = \frac{|\mathcal{P}_i \cap \mathcal{P}_j|}{|\mathcal{P}_i \cup \mathcal{P}_j|}$$

Take Figure 1(a) as an example, given the subgraph $G_1$, $S_{RP}(G_1, G_2) = \frac{2}{5}$, $S_{RP}(G_1, G_3) = \frac{1}{3}$, and $S_{RP}(G_1, G_4) = S_{RP}(G_1, G_5) = 0$. Such measurement results meet our intuitive expectation that relations play the most important role followed by the structural properties.

Path-to-Graph Index. To improve the efficiency of calculating relation path-based similarity, we build a path-to-graph index with the path identifier $p$ as the key and the set of subgraphs including the path $\pi_p$, i.e., $\mathcal{G}_p$, as the value. The whole index is denoted by $\{\mathcal{G}_p\}_{p=1}^M$ with $M$ as the number of all the distinct paths in the training data. Using the index, we can retrieve all subgraphs that contain a specific path with a complexity of $O(1)$. Then Eq. (4) can be calculated with a complexity of $O(NT)$, where $N$ is the average number of subgraphs with a path, and $T$ is the average number of paths in a subgraph.

4.3 MAML-like Meta-learner

The subgraph retriever retrieves top-$k$ similar subgraphs $\{G_j\}_{j=1}^k$ for a given $G_i$ in training data, which are used to construct the support set $\mathcal{S}_i = \{(G_j, a_i, q_i)\}_{j=1}^k$ for a learning task about the query sample $\mathcal{D}_i = \{(G_i, a_i, q_i)\}$.

For the meta-learner, the learning process consists of a meta-train step and a meta-test step. For each learning task corresponding to a sample $\mathcal{D}_i$, the meta-train step learns a task-specific learner $\theta_i'$ based on the support set $\mathcal{S}_i$:

$$\theta_i' = \theta - \alpha \nabla_{\theta} L_{BART}(f_\theta(\mathcal{S}_i)),$$

where $L_{BART}(f_\theta(\mathcal{S}_i))$ is the learner’s loss function, and $\alpha$ is the update learning rate.

To connect multiple learning tasks, the meta-test step learns the task-agnostic learner $\theta$ by the loss computed using task-specific learner $\theta_i'$:

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{D}_i \in \mathcal{D}} L_{BART}(f_{\theta_i'}(\mathcal{D}_i)),$$

where $\sum_{\mathcal{D}_i} L_{BART}(f_{\theta_i'}(\mathcal{D}_i))$ is the meta-objective, and $\beta$ is the meta learning rate. We summarize the whole procedure of the proposed DSM in Algo. 3.

5 Experimental Evaluation

We conduct extensive experiments to mainly answer the four questions: (1) Does DSM take effect in improving the KBQG performance? (2) Can the GCL-based subgraph retriever result in more effective support set for the meta-learner? (3) How do the positive sample size for contrastive learning and the support set size for the meta-learner affect DSM? (4) Can DSM benefit the QA tasks as a means of data augmentation?

5.1 Experimental Protocol

Datasets. We adopt two benchmarks, WebQuestions (WQ) and PathQuestions (PQ) (Zhou et al.,
Table 1: Data statistics. #Instances denotes the number of instances. #Entities and #Relations are the total number of entities and relations included in the dataset. #Triples represents the min/max/avg number of triplets in each instance.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Instances</th>
<th>#Entities</th>
<th>#Relations</th>
<th>#Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>WQ</td>
<td>22,989</td>
<td>25,703</td>
<td>672</td>
<td>2/99/5.8</td>
</tr>
<tr>
<td>PQ</td>
<td>9,731</td>
<td>7,250</td>
<td>378</td>
<td>2/3/2.7</td>
</tr>
</tbody>
</table>

Table 2 shows the overall evaluation results of all compared models. By the results, we summarize the following conclusions: (1) PLMs can contribute to KBQG. BART-base and BART-large show better performance than the existing three baselines, because the BART models are pretrained on the large corpus so that they are empowered with rich knowledge, while existing baselines are all trained from scratch. (2) DSM significantly outperforms BART-base and BART-large, which reflects the effectiveness of modeling the diversity of subgraphs. The baselines train subgraphs of different semantics together, which increases the learning difficulty. Alternately, we learn on and across semantics-specific tasks to capture semantics-specific and semantics-agnostic knowledge. (3) DSM shows more promising performance on the more diverse dataset WQ. We observe that WQ has more diverse subgraphs including chain, star, and triangle structures, while PQ only has the subgraphs of chain and star structures. DSM obtains only 5.03% BLEU-4 gain and 1.52% ROUGE-L gain over the best results of baselines on PQ but significantly derives 6.81% BLEU-4 gain and 7.74% ROUGE-L gain over the best results of baselines on WQ. This shows that DSM can better address the dataset with more diverse subgraphs.

5.3 Evaluation of GCL-based Retriever

We evaluate whether the proposed GCL-based retriever can result in a support set of high quality. We keep the meta-learner in DSM, and vary the retriever as 1-RP (RP is the abbreviation of Relation Path) retriever, 2-RP retriever. All-RP retriever, GED-based retriever, DGI-based retriever, and RGNCN-based retriever. The former three follow the same relation path-based similarity for generating positive samples in contrastive learning. Specifically, they enumerate the relation paths on the subgraphs and calculate the relation path similarity in Eq. (4) to retrieve top-k similar subgraphs. The differences lie in that 1-RP retriever and 2-RP retriever restrict the path length to 1 and 2 respectively, while All-RP considers all the possible paths. GED retriever retrieves top-k similar subgraphs according to the classic graph edit distance (Bunke, 1983). DGI-based retriever replaces contrastive loss in Eq. (3) with DGI loss (Velickovic et al., 2019), a “local-global contrast”. RGNCN-based retriever replaces the relation-enhanced subgraph encoder with the RGNCN encoder (Schlichtkrull et al., 2018), which considers the features of entities and relations simultaneously.
Table 2: Overall evaluation on WQ and PQ (%).

<table>
<thead>
<tr>
<th>Model</th>
<th>WQ BLEU-1</th>
<th>WQ BLEU-2</th>
<th>WQ BLEU-3</th>
<th>WQ BLEU-4</th>
<th>WQ ROUGE-L</th>
<th>PQ BLEU-1</th>
<th>PQ BLEU-2</th>
<th>PQ BLEU-3</th>
<th>PQ BLEU-4</th>
<th>PQ ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHQG+AE</td>
<td>42.35</td>
<td>29.32</td>
<td>18.43</td>
<td>9.63</td>
<td>35.72</td>
<td>45.02</td>
<td>35.86</td>
<td>28.73</td>
<td>17.86</td>
<td>63.45</td>
</tr>
<tr>
<td>G2S+AE</td>
<td>53.48</td>
<td>38.67</td>
<td>27.35</td>
<td>20.54</td>
<td>55.61</td>
<td>78.21</td>
<td>69.62</td>
<td>63.35</td>
<td>54.21</td>
<td>82.32</td>
</tr>
<tr>
<td>G2S+AE+RL</td>
<td>54.69</td>
<td>39.77</td>
<td>27.35</td>
<td>20.80</td>
<td>55.73</td>
<td>76.05</td>
<td>67.75</td>
<td>61.64</td>
<td>52.19</td>
<td>81.94</td>
</tr>
<tr>
<td>BART-base</td>
<td>56.39</td>
<td>41.05</td>
<td>29.59</td>
<td>21.46</td>
<td>56.51</td>
<td>79.59</td>
<td>70.63</td>
<td>64.30</td>
<td>55.73</td>
<td>84.54</td>
</tr>
<tr>
<td>BART-large</td>
<td>56.89</td>
<td>41.29</td>
<td>30.11</td>
<td>21.81</td>
<td>56.38</td>
<td>79.30</td>
<td>70.64</td>
<td>64.54</td>
<td>56.00</td>
<td>84.22</td>
</tr>
<tr>
<td>DSM (ours)</td>
<td>62.94</td>
<td>48.20</td>
<td>37.50</td>
<td>28.62</td>
<td>64.25</td>
<td>82.44</td>
<td>74.20</td>
<td>68.60</td>
<td>61.03</td>
<td>86.06</td>
</tr>
<tr>
<td>Performance Gain</td>
<td>6.05</td>
<td>6.91</td>
<td>7.39</td>
<td>6.81</td>
<td>7.74</td>
<td>2.85</td>
<td>3.56</td>
<td>4.06</td>
<td>5.03</td>
<td>1.52</td>
</tr>
</tbody>
</table>

Table 3: Evaluation of the GCL-based subgraph retriever and other retrievers (%).

<table>
<thead>
<tr>
<th>Retriever</th>
<th>WQ BLEU-1</th>
<th>WQ BLEU-2</th>
<th>WQ BLEU-3</th>
<th>WQ BLEU-4</th>
<th>WQ ROUGE-L</th>
<th>PQ BLEU-1</th>
<th>PQ BLEU-2</th>
<th>PQ BLEU-3</th>
<th>PQ BLEU-4</th>
<th>PQ ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-RP retriever</td>
<td>60.30</td>
<td>45.66</td>
<td>34.84</td>
<td>26.23</td>
<td>60.22</td>
<td>81.56</td>
<td>73.44</td>
<td>67.94</td>
<td>60.47</td>
<td>85.25</td>
</tr>
<tr>
<td>2-RP retriever</td>
<td>60.36</td>
<td>45.76</td>
<td>35.34</td>
<td>26.41</td>
<td>60.53</td>
<td>81.90</td>
<td>73.75</td>
<td>68.23</td>
<td>60.57</td>
<td>85.50</td>
</tr>
<tr>
<td>All-RP retriever</td>
<td>61.40</td>
<td>46.88</td>
<td>36.38</td>
<td>27.53</td>
<td>61.78</td>
<td>81.80</td>
<td>73.44</td>
<td>67.60</td>
<td>59.78</td>
<td>85.50</td>
</tr>
<tr>
<td>GED retriever</td>
<td>61.03</td>
<td>45.87</td>
<td>35.62</td>
<td>26.45</td>
<td>60.45</td>
<td>80.50</td>
<td>71.94</td>
<td>66.06</td>
<td>57.97</td>
<td>84.25</td>
</tr>
<tr>
<td>DGI-based retriever</td>
<td>61.28</td>
<td>46.57</td>
<td>36.23</td>
<td>26.58</td>
<td>62.13</td>
<td>82.10</td>
<td>73.80</td>
<td>68.25</td>
<td>60.66</td>
<td>85.80</td>
</tr>
<tr>
<td>RGCN-based retriever</td>
<td>60.87</td>
<td>45.72</td>
<td>34.57</td>
<td>25.03</td>
<td>59.24</td>
<td>80.40</td>
<td>71.56</td>
<td>65.44</td>
<td>56.84</td>
<td>84.40</td>
</tr>
<tr>
<td>GCL-based retriever</td>
<td>62.94</td>
<td>48.20</td>
<td>37.50</td>
<td>28.62</td>
<td>64.25</td>
<td>82.44</td>
<td>74.20</td>
<td>68.60</td>
<td>61.03</td>
<td>86.06</td>
</tr>
</tbody>
</table>

The bold format represents the best results over all the methods and the underline format represents the best results of baselines.

and 2-RP on the more diverse dataset WQ, while 2-RP outperforms the other two on PQ. Since PQ mostly contains the chain-style subgraphs with length 2, relation paths of length 2 could be more distinguished than other paths. On the contrary, WQ is more diverse, demanding relation paths with various lengths to express the potential structures. Thus All-RP performs better than the other two on WQ. (3) The proposed GCL-based retriever outperforms the three RP retrievers and GED retriever, because the three RP retrievers cannot find subgraphs without common relations, and the GED retriever only captures the structural knowledge but overlooks the relation semantics. (4) GCL loss function outperforms the DGI loss function. DGI aims to encode the global features of a whole subgraph into each node via the “local-global” contrast. On the contrary, GCL performs “global-global” contrast to directly compare two subgraphs, which is more suitable to the objective of the retriever. (5) RGCN encoder is worse than the relation-enhanced GNN encoder. In addition to the relation semantics, RGCN also encodes the entity information, which shows no obvious effect on determining the question semantics. Meanwhile, RGCN creates a separate parameter for each relation without considering the semantics presented by relation names, which also weakens its effect.

5.4 Sensitivity Study

We investigate how the support set size $k$ in the meta-learner and the positive sample size $m$ of GCL affect DSM. Figure 3(a) presents ROUGE-L of DSM over different support set sizes on WQ. We observe that the performance rises first then falls and reaches the top at 20. Similarly, Figure 3(b) presents ROUGE-L of DSM over different positive sample sizes for GCL on WQ. We observe the same optimal value of 20. The results indicate that more similar subgraphs might introduce additional noises. Similar trends are observed on PQ.

5.5 Positive Impacts on QA Tasks

We study whether the proposed DSM can benefit QA tasks as a means of data augmentation. We evaluate two classical KBQA models named GRAFT-Net (Sun et al., 2018a) and NSM (He et al., 2021)
on WebQSP (Yih et al., 2016), a widely-adopted KBQA dataset with 2,848 (question, answer) training instances. To evaluate the quality of the generated questions by DSM, we replace part of the (question, answer) pairs in WebQSP with the generated questions. Since the training data of WebQSP has 1,409 overlapped (question, answer) pairs with that of WQ, we can easily get their corresponding subgraphs from WQ. For easy evaluation, we replace the real questions of the 1,409 instances in WebQSP with the questions generated from the corresponding subgraphs by DSM and denote the dataset as +DSM. On this partially replaced WebQSP, we train GRAFT-Net and NSM and compare their performance with the same models trained on the partially removed dataset (-o). We also train GRAFT-Net and NSM on the datasets partially replaced by the pseudo questions generated by G2S+AE+RL and the BART-large model. We denote them as +G2S and +BART respectively.

Table 4 presents the comparison results of GRAFT-Net and NSM trained on various datasets. The results show: (1) The generated (question, answer) pairs can be viewed as a means of data augmentation for KBQA, because both GRAFT-Net and NSM trained on the datasets partially replaced by various KBQG models (i.e., +G2S, +BART, +DSM) can improve the QA performance of them trained on the partially removed dataset (i.e., -o). (2) DSM can generate much better questions than others, because the KBQA models trained on the dataset generated by DSM perform best among all the other generated datasets. (3) The generated dataset by DSM is quite close to the real data, which is supported by the comparable results between “+DSM” and “Real”.

5.6 Human Evaluation
We perform human evaluation to further verify the effectiveness of DSM. We randomly choose 100 samples $S_{100} = \{(G_j, a_j, q_j)\}_{j=1}^{100}$ from the test set of WQ dataset. Different models generate different questions for the same (subgraph, answer) pair. We evaluate the generated questions by fluency and relevance, where the former assesses whether the generated questions are readable for humans, and the latter measures the relevance between the generated question and the input (subgraph, answer) pair. We score fluency and relevance on a five-point Likert scale, with 1-point being poor and 5-point being perfect. We invite 6 annotators to score each generated question and average their scores for the proposed DSM and two baselines G2S and BART.

Table 5 presents the human evaluation results on WQ, which shows that DSM can produce more fluent and relevant questions than the other baselines, and even competes with the ground truth questions.

6 Conclusion
This work pilots studies on KBQG. We propose DSM to exploit semantic knowledge of diverse subgraphs. Instead of training on different subgraphs together, we construct semantics-specific learning tasks to reduce the learning difficulty. Specifically, we devise a GCL-based retriever to flexibly construct semantics-specific learning tasks. Besides, a MAML-like meta-learner is employed to learn on the different learning tasks, such that we can learn the semantics-specific and the semantics-agnostic knowledge shared on and across tasks. Our model shows competitive performance across the widely used benchmarks. We believe that using the MAML-like meta-learner could be inspiring for learning on datasets with high diversity.

Acknowledgments
This work is supported by National Natural Science Foundation of China (62076245, 62072460, 62172424); Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (22XNH190); National Key Research & Develop Plan(2018YFB1004401); Beijing Natural Science Foundation (4212022).
Limitations

Our model suffers from weak generalization. For example, the ground truth question is “What is the name of an attraction in Salt Lake City that has fewer than 1012563 visitors per year?” but we generate “What is the largest attraction in Salt Lake City Utah?” The model fails to generate “fewer than 1,012,563 visitors per year” because it did not see the corresponding relation “annual_visitors” during training. In another example, the ground truth question is “Who plays Jason Morgan on General Hospital as well as Cloud Strife?” but we generate “What is Jason Morgan on General Hospital?” We observe that the corresponding relation path “dubbing_performances→actor” appears in the training data, but the model still fails to generate “as well as Cloud Strife”, because the two-hop relation path is more difficult to be generalized.

References


Jiezong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. 2020. Gcc: Graph contrastive coding for graph neural network pre-training. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1150–1160.


A Notations

As shown in Table 6, the notations in this paper are described in detail.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$, $N$</td>
<td>training data and its size</td>
</tr>
<tr>
<td>$D_i=(G_i,a_i,q_i)$</td>
<td>a sample with subgraph $G_i$, answer $a_i$, and question $q_i$</td>
</tr>
<tr>
<td>$A_i$</td>
<td>the adjacent matrix of $G_i$</td>
</tr>
<tr>
<td>$S_i$</td>
<td>the support set of $D_i$</td>
</tr>
<tr>
<td>$r, c, \pi$</td>
<td>a relation, an entity, and a path</td>
</tr>
<tr>
<td>$r, h, z_i$</td>
<td>the embeddings of $r$, $c$, and $G_i$</td>
</tr>
<tr>
<td>$k, m$</td>
<td>support set size and positive sample size</td>
</tr>
<tr>
<td>$n, n_i$</td>
<td>batch size and the number of entities in $G_i$</td>
</tr>
<tr>
<td>$P_i, G_p$</td>
<td>the set of paths in $G_i$ and the set of subgraphs with $\pi_p$</td>
</tr>
<tr>
<td>$f_{\theta}, g_{\phi}$</td>
<td>the QG function and GNN encoder</td>
</tr>
</tbody>
</table>

B Experiment

B.1 Experiment Settings

For performing subgraph contrastive learning in Algo. 1, the key settings include: (1) We implement the graph encoder using the GIN framework (Xu et al., 2019) and employ the sum-style graph convolution, which sums the neighbor embeddings of a node during message passing to capture the structural properties of nodes. The layer number $L$ is set to 1, as the average number of triplets in a subgraph is only 2.7 in PQ and 5.8 in WQ. (2) For initializing the embedding of a node, the embeddings of all its connected relations are averaged. Each relation is embedded by BERT (Devlin et al., 2019). (3) For implementing the contrastive loss, we set the number of the positive samples $m$ to 20, which is the selected optimal value shown in Figure 3. (4) Following the setting of supervised contrastive learning (Khosla et al., 2020), we set the temperature parameter $\tau$ to be 0.07. In addition, we set the input feature dimension as 1024, the node representation dimension as 1024, the learning rate as 0.001, the batch size $n$ as 16, the optimizer as Adam, the patience as 15, and the maximum epochs as 100 for early stopping.

In the proposed DSM, $f$ is instantiated as BART-base. For fine-tuning BART-base in Line 7 of Algo. 3, we set the learning rate as 3e-5, batch size as 8, the patience as 15, and the maximum epochs as 50 for early stopping. BART-base has a 6-layers encoder and a 6-layers decoder. BART-large has a 12-layers encoder and a 12-layers decoder.
For fine-tuning BART-base by the meta-learner in Lines 8-14 of Algo. 3, we set the learning rate $\alpha$ for the meta-train step as $5\times 10^{-5}$, the learning rate $\beta$ for the meta-test step as $3\times 10^{-5}$, the number of tasks in a batch as 8, the meta-train steps for each task (i.e., the update steps for Line 11 in Algo. 3) as 1. We set the epochs for the meta-learner (i.e., the loop times of Lines 9-13 in Algo. 3) to 5, as the BART-base model is fine-tuned before the meta-learner training process, resulting in quick convergence.

For inference, we fine-tune the BART-base model on the support set of a new sample by 5 epochs, and then infer the top beam as the generated question of the sample.

B.2 Case Study

Due to the space limitation, we only present 13 questions generated by DSM, G2S, and BART in Table 7 on WQ. We also show the corresponding query subgraph and the support set for the top-3 cases in Figure 4. The results show that: (1) The generated question derived by our model is much closer to the ground truth question than the baselines. (2) The retrieved top-2 subgraphs in the support set are quite similar to the query subgraph in the relation semantics and the structures, so that the experience of question generation on these similar subgraphs can benefit the question generation of the query subgraph.

![Figure 4: Case study of three generated questions. The support set is built by our proposed GCL-based retriever.](image-url)
<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>G2S</th>
<th>BART</th>
<th>DSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>What actor starring on <strong>Buffy the Vampire Slayer</strong> was born in Iowa ?</td>
<td>Who plays Riley Finn in <strong>Buffy the Vampire Slayer</strong> ?</td>
<td>Who plays the character born in Iowa in <strong>Buffy the Vampire Slayer</strong> ?</td>
<td>Who plays the character born in Iowa on Buffy the Vampire Slayer ?</td>
</tr>
<tr>
<td>What are the common beliefs of the religions that believe in the end times ?</td>
<td>What do the people who believes in end time ?</td>
<td>What did the person who believed in the end time practice ?</td>
<td>What are some of the beliefs of the religion that believes in the end time theory practiced by Abrahamic people ?</td>
</tr>
<tr>
<td>What time zone would I be in if I was in the US State whose capital is Jefferson city ?</td>
<td>What time zone is the state whose capital is Jefferson city ?</td>
<td>What is the time zone where Jefferson city is the capital ?</td>
<td>What time zone is used in the US State whose capital is Jefferson city ?</td>
</tr>
<tr>
<td>What country export to Sudan and has Giza necropolis ?</td>
<td>What country exports to Sudan and Sudan ?</td>
<td>What country that exports to Sudan is Giza necropolis located in ?</td>
<td>What country exports to Sudan and is home to the Giza necropolis ?</td>
</tr>
<tr>
<td>What movie produced by Brad Lewis did Alyson Stoner starred in ?</td>
<td>What movie was produced by Brad Lewis ?</td>
<td>What movie produced by Brad Lewis was Alyson Stoner in ?</td>
<td>What movie produced by Brad Lewis did Alyson Stoner play in ?</td>
</tr>
<tr>
<td>Which english language shows did Henry Winkler produce ?</td>
<td>What show produced by Henry Winkler has Henry Winkler as an actor ?</td>
<td>What english speaking shows did Henry Winkler produce ?</td>
<td>What english speaking shows did Henry Winkler produce ?</td>
</tr>
<tr>
<td>Which type of monarchy does Japan have that is similar as the Kingdom of Prussia ?</td>
<td>What type of government is used in Japan and Japan ?</td>
<td>What type of government can be found in both Japan and the Kingdom of Prussia ?</td>
<td>What type of government is used in both Japan and the Kingdom of Prussia ?</td>
</tr>
<tr>
<td>What actress played Rose Loomis and has ties to John F. Kennedy ?</td>
<td>Which actor who played the character Rose Loomis ?</td>
<td>What actress who portrayed Rose Loomis was John F. Kennedy dating ?</td>
<td>Which actress played Rose Loomis and also dated John F. Kennedy ?</td>
</tr>
<tr>
<td>in the Tortall universe what language do native American Indians speak ?</td>
<td>What language is spoken in the &lt;unk&gt; and the fictional universe the Tortall universe ?</td>
<td>What language do native American Indians speak ?</td>
<td>What language, found in the fictional Tortall universe, do native American Indians speak ?</td>
</tr>
<tr>
<td>What person born in Batlesville was the first leader of the AFL ?</td>
<td>Who was born in &lt;unk&gt; and was born in &lt;unk&gt; ?</td>
<td>Who was born in Bartlesville and was a person of the AFL ?</td>
<td>Who was born in Bartlesville and was a member of the AFL first team ?</td>
</tr>
<tr>
<td>Who participated in the third joint debate at Jonesboro and influenced Walt Whitman’s poetry ?</td>
<td>Who was the speaker of the speaker at &lt;unk&gt; ?</td>
<td>Which speaker participated in the third joint debate at Jonesboro ?</td>
<td>Which speaker featured in the third joint debate at Jonesboro influenced Walt Whitman ?</td>
</tr>
<tr>
<td>What country is home to Nova Roma and borders Bolivia ?</td>
<td>What country bordering Bolivia and Nova Roma ?</td>
<td>What country borders Bolivia and Nova Roma ?</td>
<td>What country borders Bolivia and is home to Nova Roma ?</td>
</tr>
<tr>
<td>What movie featuring Rihanna was released last ?</td>
<td>What is the earliest released film that Rihanna starred in ?</td>
<td>What is the latest released film that Rihanna starred in ?</td>
<td>What is the latest film that Rihanna has been in that was released last in 2012 ?</td>
</tr>
</tbody>
</table>

"<unk>" represents a word that does not appear in the vocabulary.